Adaptive Clinical Trials

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Center for Devices and Radiologic Health

November 8, 2010

1

Preliminaries

Conflicts of Interest / Disclaimers

2

Conflicts of Interest

- Commercially available software for sequential clinical trials
  - S+SeqTrial (Emerson)
    - Planned release as freely available module in R in January, 2011
      - www.RCTdesign.org
  - PEST (Whitehead)
  - EaSt (Mehta)
  - SAS (I provided some limited advice)

3

Personal Defects

- Physical
- Personality
  - In the perjorative sense of the words
    - Then:
      - A bent twig
    - Now:
      - Old and male
      - University professor

4

A More Descriptive Title

- All my talks

The Use of Statistics to Answer

Scientific Questions

(Confessions of a Former Statistician)

5

A More Descriptive Title

- All my talks

The Use of Statistics to Answer

Scientific Questions

(Confessions of a Former Statistician)

- Clinical Trial Design talks

The Use of Statistics to Answer

Scientific Questions Ethically and Efficiently

(Confessions of a Former Statistician)

6

Overview of Clinical Trial Design

Science and Statistics

Where am I going?

In the real world, clinical trial design must consider

- scientific theory
- statistical theory
- logistical issues
- game theory

I make an argument (plea?) for clinical trial design to consider science first, then statistics

(Game theory is a necessary evil)

7

Clinical Trials

- Experimentation in human volunteers
- Investigates a new treatment/preventive agent
  - Safety:
    - Are there adverse effects that clearly outweigh any potential benefit?
  - Efficacy:
    - Can the treatment alter the disease process in a beneficial way?
  - Effectiveness:
    - Would adoption of the treatment as a standard affect morbidity / mortality in the population?

8

Carrying Coals to Newcastle

- Wiley Act (1906)
  - Labeling
- Food, Drug, and Cosmetics Act of 1938
  - Safety
- Kefauver – Harris Amendment (1962)
  - Efficacy / effectiveness
    - " [If] there is a lack of substantial evidence that the drug will have the effect ... shall issue an order refusing to approve the application. "
    - "...The term 'substantial evidence' means evidence consisting of **adequate and well-controlled investigations, including clinical investigations**, by experts qualified by scientific training"
- FDA Amendments Act (2007)
  - Registration of RCTs, Pediatrics, Risk Evaluation and Mitigation Strategies (REMS)

9

Medical Devices

- Medical Devices Regulation Act of 1976
  - Class I: General controls for lowest risk
  - Class II: Special controls for medium risk - 510(k)
  - Class III: Pre marketing approval (PMA) for highest risk
    - "…**valid scientific evidence** for the purpose of determining the safety or effectiveness of a particular device … adequate to support a determination that there is reasonable assurance that the device is safe and effective for its conditions of use…"
    - "Valid scientific evidence is evidence from well-controlled investigations, partially controlled studies, studies and objective trials without matched controls, well-documented case histories conducted by qualified experts, and reports of significant human experience with a marketed device, **from which it can fairly and responsibly be concluded by qualified experts that there is reasonable assurance of the safety and effectiveness**…"
- Safe Medical Devices Act of 1990
  - Tightened requirements for Class 3 devices

10

The Problem of Clinical Trial Design

Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy, *Anna Karenina, 1873-77*

11

The Problem of Clinical Trial Design

Unbiased clinical trials are all alike; every biased clinical trial is biased in its own way.

12

Clinical Trial Design

- Finding an approach that best addresses the often competing goals: Science, Ethics, Efficiency
  - Basic scientists: focus on mechanisms
  - Clinical scientists: focus on overall patient health
  - Ethical: focus on patients on trial, future patients
  - Economic: focus on profits and/or costs
  - Governmental: focus on safety of public: treatment safety, efficacy, marketing claims
  - Statistical: focus on questions answered precisely
  - Operational: focus on feasibility of mounting trial

13

Statistical Planning

- Satisfy collaborators as much as possible
- Discriminate between relevant scientific hypotheses
  - Scientific and statistical credibility
- Protect economic interests of sponsor
  - Efficient designs
  - Economically important estimates
- Protect interests of patients on trial
  - Stop if unsafe or unethical
  - Stop when credible decision can be made
- Promote rapid discovery of new beneficial treatments

14

Classical Fixed Sample Designs

- Design stage:
  - Choose a sample size
- Conduct stage:
  - Recruit subjects, gather all the data
- Analysis stage:
  - When all data available, analyze and report

15

Statistical Sampling Plan

- Ethical and efficiency concerns are addressed through sequential sampling
- During the conduct of the study, data are analyzed at periodic intervals and reviewed by the DMC
- Using interim estimates of treatment effect
  - Decide whether to continue the trial
    - Stop if scientifically relevant and statistically credible results have been obtained
  - If continuing, decide on any modifications to
    - sampling scheme
    - scientific / statistical hypotheses and/or

16

Ultimate Goal

- Modify the sample size accrued so that minimal number of subjects treated when
  - new treatment is harmful,
  - new treatment is minimally effective, or
  - new treatment is extremely effective
- Only proceed to maximal sample size when
  - not yet certain of treatment benefit, or
  - potential remains that results of clinical trial will eventually lead to modifying standard practice

17

Group Sequential Designs

- Design stage:
  - Choose an interim monitoring plan
  - Choose a maximal stopping time
    - Statistical information, sample size, calendar time
- Conduct stage:
  - Recruit subjects, gather data in groups
  - After each group, analyze for DMC
    - DMC recommends termination or continuation
- Analysis stage:
  - When study stops, analyze and report

18

Group Sequential Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - Can be randomly determined if independent of effect
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- Compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue

19

Statistical Design Issues

- Under what conditions should we use fewer subjects?
  - Ethical treatment of patients
  - Efficient use of resources (time, money, patients)
  - Scientifically meaningful results
  - Statistically credible results
  - Minimal number of subjects for regulatory agencies
- How do we control false positive rate?
  - Repeated analysis of accruing data involves multiple comparisons

20

Distinctions without Differences

- Sequential sampling plans
  - Group sequential stopping rules
  - Error spending functions
  - Conditional / predictive power
  - Bayesian posterior probabilities
- Statistical treatment of hypotheses
  - Superiority / Inferiority / Futility
  - Two-sided tests / bioequivalence

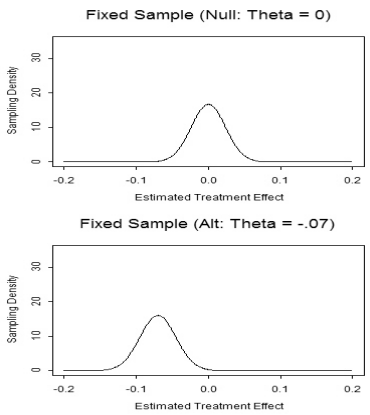
21

Major Statistical Issue

- Frequentist operating characteristics are based on the sampling distribution
- Stopping rules do affect the sampling distribution of the usual statistics
  - MLEs are not normally distributed
  - Z scores are not standard normal under the null
    - (1.96 is irrelevant)
  - The null distribution of fixed sample P values is not uniform
    - (They are not true P values)
- Bayesian operating characteristics are based on the sampling distribution and a prior distribution
  - Is stopping rule relevant?

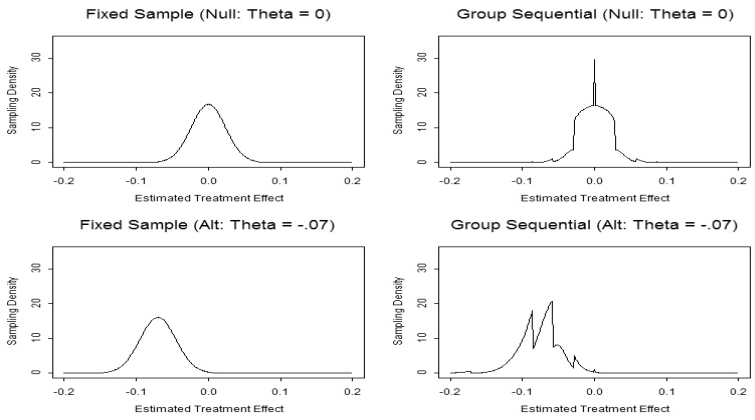
22

Sampling Distribution of MLE

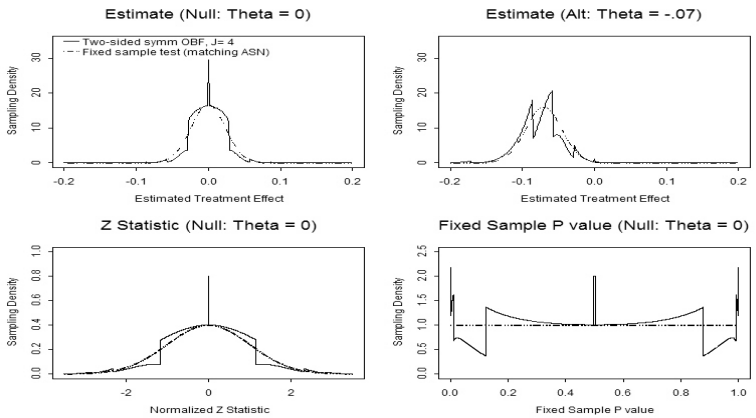


23

Sampling Distribution of MLE



Sampling Distributions



So What?

- Demystification: All we are doing is statistics
  - Planning a study
  - Gathering data
  - Analyzing it

26

Sequential Studies

- Demystification: All we are doing is statistics
  - Planning a study
    - Added dimension of considering time required
  - Gathering data
    - Sequential sampling allows early termination
  - Analyzing it
    - The same old inferential techniques
    - The same old statistics
    - But new sampling distribution

27

Familiarity and Contempt

- For any known stopping rule, however, we can compute the correct sampling distribution with specialized software
  - Standalone programs
    - PEST (some integration with SAS)
    - EaSt
  - Within statistical packages
    - S-Plus S+SeqTrial (soon R)
    - SAS PROC SEQDESIGN

28

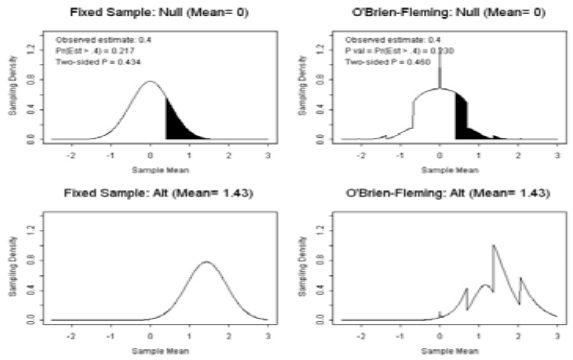
Familiarity and Contempt

- From the computed sampling distributions we then compute
  - Bias adjusted estimates
  - Correct (adjusted) confidence intervals
  - Correct (adjusted) P values
- Candidate designs can then be compared with respect to their operating characteristics

29

Example: P Value

- Null sampling density tail



30

Evaluation of Designs

- Define candidate design constraining two operating characteristics
  - Type I error, power at design alternative
  - Type I error, maximal sample size
- Evaluate other operating characteristics
  - Sample size requirements
  - Power curve
  - Inference to be reported upon termination
  - (Probability of a reversed decision)
- Modify design
- Iterate

31

But What If …?

- Possible motivations for adaptive designs
  - Changing conditions in medical environment
    - Approval / withdrawal of competing / ancillary treatments
    - Diagnostic procedures
  - New knowledge from other trials about similar treatments
  - Evidence from ongoing trial
    - Toxicity profile (therapeutic index)
    - Interim estimates of primary efficacy / effectiveness endpoint
      - Overall
      - Within subgroups
    - Interim alternative analyses of primary endpoints
    - Interim estimates of secondary efficacy / effectiveness endpoints

32

Adaptive Sampling Plans

- At each interim analysis, possibly modify
  - Maximal statistical information
  - Schedule of analyses
  - Conditions for early stopping
  - Randomization ratios
  - Statistical criteria for credible evidence
  - Scientific and statistical hypotheses of interest

33

Adaptive Sampling: Examples

- Response adaptive modification of sample size
  - Proschan & Hunsberger (1995); Cui, Hung, & Wang (1999)
- Response adaptive randomization
  - Play the winner (Zelen, 1979)
- Adaptive enrichment of promising subgroups
  - Wang, Hung & O'Neill (2009)
- Adaptive modification of endpoints, eligibility, dose, …
  - Bauer & Köhne (1994); LD Fisher (1998)

34

Adaptive Sampling: Issues

- How do the newer adaptive approaches relate to the constraint of human experimentation and scientific method?
- Effect of adaptive sampling on trial ethics and efficiency
  - Avoiding unnecessarily exposing subjects to inferior treatments
  - Avoiding unnecessarily inflating the costs (time / money) of RCT
- Effect of adaptive sampling on scientific interpretation
  - Exploratory vs confirmatory clinical trials
- Effect of adaptive sampling on statistical credibility
  - Control of type I error in frequentist analyses
  - Promoting predictive value of "positive" trial results

35

Adaptive Sample Size Determination

- Design stage:
  - Choose an interim monitoring plan
  - Choose an adaptive rule for maximal sample size
- Conduct stage:
  - Recruit subjects, gather data in groups
  - After each group, analyze for DMC
    - DMC recommends termination or continuation
  - After penultimate group, determine final N
- Analysis stage:
  - When study stops, analyze and report

36

Adaptive Sample Size Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - $N_1 \ldots N_{J-1}$ can be randomly determined indep of effect
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- At $N_1 \ldots N_{J-1}$ compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue
    - At N_{J-1} determine N_J according to value of T_j

37

When Stopping Rules Not Pre-specified

- Methods to control the type I error have been described for fully adaptive designs
  - Most popular: Preserve conditional error function from some fixed sample or group sequential design
- Methods to compute bias adjusted estimates and confidence intervals not yet well-developed

38

Scientific Design Issues

- Under what conditions should we alter our hypotheses?
  - Definition of a "Treatment Indication"
    - Disease
    - Patient population
    - Definition of treatment (dose, frequency, ancillary treatments, …)
    - Desired outcome
- How do we control false positive rate?
  - Multiple comparisons across subgroups, endpoints
- How do we provide inference for Evidence Based Medicine?

39

Adaptive Sampling: Issues

"Make new friends, but don't forget the old.
One is silver and the other is gold"

- Children's song

40

CDER/CBER Guidance on Adaptive Designs

- Recommendations for use of adaptive designs in confirmatory studies
  - Fully pre-specified sampling plans
  - Use well understood designs
    - Fixed sample
    - Group sequential plans
    - Blinded adaptation
  - For the time-being, avoid less understood designs
    - Adaptation based on unblinded data
- Adaptive designs may be of use in early phase clinical trials

41

Perpetual Motion Machines

- A major goal of this course is to discern the hyperbole in much of the recent statistical literature
  - The need for adaptive clinical trial designs
    - What can they do that more traditional designs do not?
  - The efficiency of adaptive clinical trial designs
    - Has this been demonstrated?
    - Are the statistical approaches based on sound statistical foundations?
  - "Self-designing" clinical trials
    - When are they in keeping with our scientific goals?

42

Important Distinctions in this Course

- What aspects of the RCT are modified?
  - *Statistical*: Modify only the sample size to be accrued
  - *Scientific*: Possibly modify the hypotheses related to patient population, treatment, outcomes
- Are all planned modifications described at design?
  - *"Prespecified adaptive rules":* Investigators describe
    - Conditions under which trial will be modified and
    - What those modification will consist of
  - *"Fully adaptive":* At each analysis, investigators are free to use current data to modify future conduct of the study
- When do the modifications take place?
  - *Immediately:* Potential to stop study at current analysis
  - *Future stages of the RCT:* Change plans for next analysis

43

Statistics and Science

- Statistics is about science
  - Science in the broadest sense of the word
- Science is about proving things to people
  - Science is necessarily adversarial
    - Competing hypotheses to explain the real world
  - Proof relies on willingness of the audience to believe it
  - Science is a process of successive studies
- Game theory: Accounting for conflicts of interest
  - Financial
  - Academic / scientific

44

Science vs Statistics

- Recognizing the difference between
  - The parameter space
    - What is the true scientific relationship?
  - The sample space
    - What data will you / did you gather?

45

"Parameter" vs "Sample" Relationships

- The true scientific relationship ("parameter space")
  - Summary measures of the effect in population
    - Means, medians, geometric means, proportions…
- Scientific "sample space" scales:
  - Estimates attempting to assess scientific importance
    - Point estimate is a statistic estimating a "parameter"
    - Interval estimates
      - CI describes the values in the "parameter space" that are consistent with the data observed (the "sample space")
- Purely statistical "sample space" scales
  - The precision with which you know the true effect
    - Power, predictive (conditional) power, P values, posterior probabilities

46

Bottom Line

You better think (think)

about what you're

trying to do…

-Aretha Franklin, "Think"

47

Experimentation Directed Toward Adopting New Treatments

Phases of Investigation

Where am I going?

The investigation of new treatments, preventive strategies, and diagnostic procedures typically progresses through several phases.

The use of adaptive clinical trial design will need to consider the stage of investigation and purpose of the clinical trial

– Some adaptations are antithetical to confirmatory trial

– Some adaptive approaches may better control exploration

48

Overall Goal

- "Drug discovery"
  - More generally
    - a therapy / preventive strategy or diagnostic / prognostic procedure
    - for some disease
    - in some population of patients
- A series of experiments to establish
  - Safety of investigations / dose
  - Safety of therapy
  - Measures of efficacy
    - Treatment, population, and outcomes
  - Confirmation of efficacy
  - Confirmation of effectiveness

49

Phases of Investigation

- Series of studies support adoption of new treatment
  - Preclinical
    - Epidemiology including risk factors
    - Basic science:
      - Biochemistry
      - Physiologic mechanisms
      - Physics / engineering
    - Animal experiments: Toxicology / safety
  - Clinical
    - Phase I: Initial safety / dose finding
    - Phase II: Preliminary efficacy / further safety
    - Phase III: Confirmatory efficacy / effectiveness
  - Approval of indication
    - (Phase IV: Post-marketing surveillance, REMS)

50

The Enemy

"Let's start at the very beginning, a very good place to start…"

- Maria von Trapp

(as quoted by Rodgers and Hammerstein)

51

First

- Where do we want to be?
  - Find a new treatment that improves health of individuals
  - Find a new treatment that improves health of the population

52

Treatment "Indication"

- Disease
  - Therapy: Putative cause vs signs / symptoms
    - May involve method of diagnosis, response to therapies
  - Prevention / Diagnosis: Risk classification
- Population
  - Therapy: Restrict by risk of AEs or actual prior experience
  - Prevention / Diagnosis: Restrict by contraindications
- Treatment or treatment strategy
  - Formulation, administration, dose, frequency, duration, ancillary therapies
- Outcome
  - Clinical vs surrogate; timeframe; method of measurement

53

Experimental Results

- Start with the science
  - No "experiment" is ethical if it cannot answer a relevant question
- Hypotheses are the spectrum of possible results
  - An "experiment" discriminates among the possible hypotheses
  - The Scientist Game

54

Scientific Method

- Planned experiment includes protocol specified in advance, including
  - Overall goal
  - Specific aims
  - Materials: Patients, treatments
  - Methods: Administration, monitoring, outcomes
  - Methods: Statistical analysis plan
    - Sampling plan
    - Statistical models for analysis
    - Planned interpretation of spectrum of results

55

Specific Aim

- One of a series of studies used to support adoption of a new standard of treatment
  - Phase I: Initial safety / dose finding
  - Phase II: Preliminary efficacy / further safety
  - Phase III: "Registrational trials"
    - Therapeutics: Establish effectiveness
    - Prevention: Establish efficacy
    - Diagnostics: Establish accuracy
  - Phase IV:
    - Therapeutics: Post-marketing surveillance
    - Prevention: Effectiveness
    - Diagnostics: Impact on outcomes

56

Phase III Confirmatory Trials

- The major goal of a "registrational trial" is to confirm a result observed in some early phase study
- Rigorous science: Well defined confirmatory studies
  - Eligibility criteria
  - Comparability of groups through randomization
  - Clearly defined treatment strategy
  - Clearly defined clinical outcomes (methods, timing, etc.)
  - Unbiased ascertainment of outcomes (blinding)
  - Prespecified primary analysis
    - Population analyzed as randomized
    - Summary measure of distribution (mean, proportion, etc.)
    - Adjustment for covariates

57

Why Emphasize Confirmatory Trials?

"When you go looking for something specific, your chances of finding it are very bad, because of all the things in the world, you're only looking for one of them.

"When you go looking for anything at all, your chances of finding it are very good, because of all the things in the world, you're sure to find some of them."

- Darryl Zero in "The Zero Effect"

58

Why Emphasize Confirmatory Trials?

"When you go looking for something specific, your chances of finding [a spurious association by chance] are very bad, because of all the things in the world, you're only looking for one of them.

"When you go looking for anything at all, your chances of finding [a spurious association by chance] are very good, because of all the things in the world, you're sure to find some of them."

59

Real-life Examples

- Effects of arrhythmias post MI on survival
  - Observational studies: high risk for death
  - CAST: Specific anti-arrhythmics have higher mortality
- Effects of beta-carotene on lung CA and survival
  - Observational studies: high dietary beta carotene has lower cancer incidence and longer survival
  - CARET: beta carotene supplementation in smokers leads to higher lung CA incidence and lower survival
- Effects of hormone therapy on cardiac events
  - Observational studies: HT has lower cardiac morbidity and mortality
  - WHI: HT in post menopausal women leads to higher cardiac mortality

60

Multiple Comparisons in Biomedicine

- Observational studies
  - Observe many outcomes
  - Observe many exposures
  - Perform many alternative analyses
    - Summary of outcome distribution, adjustment for covariates
  - Consequently: Many apparent associations
    - May be type I errors
    - But even when valid, may be poorly understood due to confounding
- Interventional experiments
  - Exploratory analyses ("Drug discovery")
    - Modification of analysis methods
    - Multiple endpoints
    - Restriction to subgroups

61

Mathematical Basis

- The multiple comparison problem is traced to a well known fact of probability

$$Pr(A \text{ or } B) \geq Pr(A)$$

$$Pr(A \text{ or } B) \geq Pr(B)$$

62

Statistics and Game Theory

- Multiple comparison issues
  - Type I error for each endpoint – subgroup combination
    - In absence of treatment effect, will still decide a benefit exists with probability, say, .025 in each such combination
- Multiple endpoints and subgroups increase the chance of deciding an ineffective treatment should be adopted
  - This problem exists with either frequentist or Bayesian criteria for evidence
  - The actual inflation of the type I error depends
    - the number of multiple comparisons, and
    - the correlation between the endpoints

63

Type I Error Inflation: Endpoints, Subgroups

- Experiment-wise error rate from multiple level .05 tests
  - Alternative summary measures are positively correlated
  - Alternative clinical endpoints are usually positively correlated
  - Subgroups defined by the same variable are independent

Number Compared	Worst Case	Correlation 0.00	0.30	0.50	0.75	0.90
1	.050	.050	.050	.050	.050	.050
2	.100	.098	.095	.090	.081	.070
3	.150	.143	.137	.126	.104	.084
5	.250	.226	.208	.184	.138	.101
10	.500	.401	.353	.284	.193	.127
20	1.000	.642	.540	.420	.258	.154
50	1.000	.923	.806	.624	.353	.193

64

Type I Error Inflation: Summary Measures

- Example: Type I error with **normal** data
  - Consider six different summary measures

Any single test:	**0.050**
Mean, geometric mean	**0.057**
Mean, Wilcoxon	**0.061**
Mean, geom mean, Wilcoxon	**0.066**
Above plus median	**0.085**
Above plus Pr (Y > 1 sd)	**0.127**
Above plus Pr (Y > 1.645 sd)	**0.169**

65

Type I Error Inflation: Summary Measures

- Example: Type I error with **lognormal** data
  - Consider six different summary measures

Any single test:	**0.050**
Mean, geometric mean	**0.074**
Mean, Wilcoxon	**0.077**
Mean, geom mean, Wilcoxon	**0.082**
Above plus median	**0.107**
Above plus Pr (Y > 1 sd)	**0.152**
Above plus Pr (Y > 1.645 sd)	**0.192**

66

Phase III Clinical Trials

- Establishment of efficacy / effectiveness
  - Goals:
    - Obtain measure of treatment's efficacy on disease process
    - Incidence of major adverse effects
    - Therapeutic index
    - Modify clinical practice (obtain regulatory approval)
  - Methods
    - Relatively large number of participants from true target population (almost)
    - Clinically relevant outcome

67

Efficacy: A Moving Target

- Definition of efficacy can vary widely according to choice of endpoint and magnitude of importance
  - Basic science
    - Does treatment have any effect on the pathway
  - Clinical science
    - Does treatment have a sufficiently large effect on a clinically relevant endpoint

68

Effectiveness: A Moving Target

- A treatment is “effective” if its introduction improves health in the population
- A treatment can be both efficacious and ineffective depending on factors of clinical trials
  - Target population
  - Control treatment
  - Intervention
  - Measurement of outcome(s)
  - Summary measure of outcome distribution

69

Target Population

- Efficacy and effectiveness study populations may differ with respect to
  - Properly diagnosed disease
  - Subgroups with more (less) severe disease
  - Prior experience with treatment(s)
    - Randomized withdrawal
  - Ancillary treatments
  - Different risk factors

70

Control Treatment

- Efficacy and effectiveness study comparators may differ with respect to
  - Use of existing alternative treatments
  - Allowed ancillary treatments

71

Intervention

- Efficacy and effectiveness studies may differ with respect to
  - Dose
  - Administration
  - Duration
  - Training
  - Quality control

72

Measurement of Outcome

- Efficacy and effectiveness studies may differ with respect to
  - Importance of measurement: clinical vs subclinical
  - Timing of measurement: short vs long term
  - Summary measure of distribution: mean vs relevant threshold

73

Ex: ITD in OOHCA

- Impedance Threshold Device in Pre-Hospital Cardiac Arrest
- Efficacy trial might consider
  - Modified intent to treat in VT/VF
  - Administration of device within 4 minutes of call to 911
  - Exclude use of filter on ventilation mask
  - Strict protocol for CPR delivery
  - Strict protocol for monitoring of patient in ED / Hospital
  - Outcome is survival to ED or 24 hours
- Effectiveness populations might include
  - All patients treated by EMS for OOHCA
  - Ancillary treatments per current standards
  - Outcome is neurologically intact survival at 6 months

74

Ex: Screening for Lung Cancer

- RCT of spiral CT versus chest X-ray (cf: NLST)
- Efficacy trial might consider
  - High risk, asymptomatic smokers
  - Three annual screens by highly trained radiologists
  - Protocol defined follow-up screening / treatment
  - Outcome is cancer diagnosis or mortality due to lung cancer
- Effectiveness trial might consider
  - Broader entry criteria
  - Annual screening by community radiologists until end of study
  - Follow-up per local medical standards
  - Outcome is all cause mortality

75

Which: Efficacy or Effectiveness

- Factors leading to efficacy trials
  - “Knowledge is good”
  - As pilot studies before prevention studies
  - Inability to perform experiment under realistic conditions
- Factors leading to effectiveness trials
  - Serious conditions
    - Patients generally want to get better
  - Short therapeutic window for treatment
  - Waiver of informed consent
    - Do not withhold beneficial treatments in order to establish mechanisms
  - High cost of clinical trials (time, people, $$)

76

Phase III Clinical Trials: Settings

- Phase III: Common scenarios
  - Establish efficacy / effectiveness of new treatment
    - superiority over no intervention
    - superiority over existing treatment
  - Establish equivalence with current treatment
    - Two-sided equivalence: bioequivalence
      - establish response not markedly higher or lower
    - One-sided equivalence: noninferiority
      - establish treatment not markedly worse
      - perhaps superior on secondary endpoint
  - Establish harm of existing treatment

77

Potential Deleterious Impact of Adaptation

- Adaptive modification of scientific hypotheses may destroy the scientific and regulatory relevance of the trial
  - Modification of patient population, treatment, outcomes will change the hypothesized indication
- Impact on evidence based medicine
  - Physicians need to judge the magnitude of treatment effect in order to choose among alternatives for individual patients
  - Even with control of the experimentwise type I error, quantification of treatment effect will be biased with adaptation
    - Sampling distribution for the “winning” indication will depend on the true effects of the alternative indications that were dropped
    - There will undoubtedly be “regression to the true mean” on the subsequently gathered data
    - There is “random high bias” in the previously gathered data

78

Need for Exploratory Science

- Before we can do a large scale, confirmatory Phase III trial, we must have
  - A hypothesized treatment indication to confirm
    - Disease
    - Patient population
    - Treatment strategy
    - Outcome
  - Comfort with the safety / ethics of human experimentation
- In “drug discovery”, in particular, we will not have much experience with the intervention

79

Phase II Clinical Trials

- Preliminary evidence of efficacy
  - Goals:
    - Screening for any evidence of treatment efficacy
    - Incidence of major adverse effects
    - Decide if worth studying in larger samples
      - Gain information about best chance to establish efficacy
        - » Choose population, treatment, outcomes
  - Methods
    - Relatively small number of participants
    - Participants closer to true target population
    - Outcome often a surrogate
    - Sometimes no comparison group (especially in cancer)

80

Screening Studies as Diagnostic Tests

- Clinical testing of a new treatment, preventive agent, or diagnostic method is analogous to using laboratory or clinical tests to diagnose a disease
  - Goal is to find a procedure that identifies truly beneficial interventions
- Not surprisingly, the issues that arise when screening for disease apply to clinical trials
  - Predictive value of a positive test is best when prevalence is high
  - Use screening trials to increase prevalence of beneficial treatments

81

Preliminary Studies in Screening

- Two general approaches to studying new treatments
  - Study every treatment in a large definitive experiment
    - Only do Phase III studies
      - Level of significance 0.025, high power
    - (Ignore, for now, the safety / ethics of this)
  - Perform small screening trials, with confirmatory trials of promising treatments passing early tests
    - Phase II studies
      - Level of significance, power (sample size) to be determined
    - Confirmatory
      - Level of significance 0.025, high power

82

Scenario 1: Only Phase III

- Only large trials using 1,000,000 subjects
  - 10% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 1,000 subjects provide 97.5% power ➔ 1,000 RCT
    - 511 subjects provide 80.0% power ➔ 1,958 RCT
    - 250 subjects provide 50.0% power ➔ 4,000 RCT
  - Results
    - N= 1,000: 98 effective / 23 ineffective (PV+ = .81)
    - N= 511: 157 effective / 44 ineffective (PV+ = .78)
    - N= 250: 200 effective / 90 ineffective (PV+ = .69)

83

Scenario 2a: Screening Phase II

- Use 700,000 subjects in Phase II studies
  - 10% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 56 subjects provide 15% power ➔ 12,611 RCT
  - Results
    - N= 56: 189 effective / 284 ineffective (PV+ = .40)

84

Scenario 2a: Confirmatory Phase III

- Use 300,000 subjects in confirmatory Phase III studies
  - 40% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 634 subjects provide 87.7% power ➔ 473 RCT
  - Results
    - N= 634: 166 effective / 7 ineffective (PV+ = .96)

85

Scenario 2a: Comparison

- Scenario 1: Only large trials
  - Test 1 - 4K treatments (10% eff, α = 0.025)
    - Power 0.975: 98 effective / 23 ineffective (PV+ = .81)
    - Power 0.800: 157 effective / 44 ineffective (PV+ = .78)
    - Power 0.500: 200 effective / 90 ineffective (PV+ = .69)
- Scenario 2a: Use of pilot studies (70% N)
  - Screen 12,611 treatments (10% eff, α = 0.025)
    - Power 0.150: 189 effective / 284 ineffective (PV+ = .40)
  - Test 473 treatments (40% eff, α = 0.025)
    - Power 0.877: 166 effective / 7 ineffective (PV+ = .96)

86

Scenario 2b: Screening Phase II

- Use 700,000 subjects in Phase II studies
  - 10% of drugs being investigated truly work
  - Level of significance .10
  - Sample size / power
    - 350 subjects provide 85% power ➔ 2,082 RCT
  - Results
    - N= 350: 170 effective / 180 ineffective (PV+ = .49)

87

Scenario 2b: Confirmatory Phase III

- Use 300,000 subjects in confirmatory Phase III studies
  - 49% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 856 subjects provide 95% power ➔ 350 RCT
  - Results
    - N= 856: 162 effective / 5 ineffective (PV+ = .97)

88

Scenario 2b: Comparison

- Scenario 1: Only large trials
  - Test 1 – 4K treatments (10% eff, α = 0.025)
    - Power 0.975: 98 effective / 23 ineffective (PV+ = .81)
    - Power 0.800: 157 effective / 44 ineffective (PV+ = .78)
    - Power 0.500: 200 effective / 90 ineffective (PV+ = .69)
- Scenario 2b: Use of pilot studies (70% N)
  - Screen 2,082 treatments (10% eff, α = 0.10)
    - Power 0.850: 170 effective / 180 ineffective (PV+ = .49)
  - Test 350 treatments (49% eff, α = 0.025)
    - Power 0.952: 162 effective / 5 ineffective (PV+ = .97)

89

Burden of Larger Phase II Studies?

- Sequential sampling
  - Aggressive early stopping for futility
    - Greatest efficiency
  - Conservative early stopping for efficacy
    - Burden of proof, other endpoints
- Average sample size requirements
  - Only large studies: 58.5% of fixed sample
  - Pilot scenario 1 : 56.0%
  - Pilot scenario 2 : 61.0%

90

Screening Phase II: Bottom Line

- Pilot studies increase the predictive value of a positive study while using the same number of subjects.
  - Screening parameters can be optimized
    - Proportion of subjects in Phase II vs Phase III
    - Type I error at Phase II
    - Power at Phase II
- But need to consider the number of RCT and the prevalence of effective treatments
  - Will we have same prevalence of “good” ideas when we have 1,000 RCT vs 12,611 RCT?

91

Phase II Clinical Trials: Methods

- As typically implemented, the screening role of Phase II RCT is somewhat attenuated
  - Disease definition may be restrictive due to desires for
    - Efficacy: to demonstrate proof of concept
    - Safety / ethics: Caution with the unknown treatment in extremely serious disease (e.g., cancer)
  - Participants may not be true target population
    - Heavier trial burden in early trials
    - Unknown impact of concomitant disease
  - Outcome often a surrogate
    - Reduce costs / duration of RCT
    - Plausibility of effect on clinical outcome

92

Phase II Clinical Trials: Results

- The design of confirmatory Phase III trials often uses exploratory analyses from Phase II trials
  - More knowledge is now available
    - Lessened safety / ethical concerns
    - Less need for heavy trial burden
  - Exploratory subgroup analyses
    - More / less severe disease
    - Restrictions based on concomitant disease
    - Modifications of treatment intensity, ancillary treatments
    - Try for optimal definition of clinical outcome, summary measures, statistical tests

93

Potential Beneficial Impact of Adaptation

- The "data dredging" that frequently occurs between phase II and phase III will tend to inflate the experiment-wise type I error
- Some of the adaptive clinical trial approaches control experimentwise type I error while attempt to allow
  - Modification of patient population
  - Modification of treatment
  - Modification of endpoints
  - Modification of statistical analysis models
- To the extent that the added control encourages adherence to the desirable operating characteristics, the adaptive methods can be extremely advantageous

94

Seamless Phase II / III Clinical Trials

- Much recent interest in moving quickly from Phase II to Phase III studies
  - Plan a Phase III study
  - Incorporate early analysis of data to assess whether complete study
    - Early analysis may be based on different endpoint
  - Comments
    - Major gain is related to timeline of accrual
    - Only feasible if no changes based on early results
      - Same eligibility, treatment, measurement of outcomes
    - Blurs role of screening and confirmation
      - Should early phase data be included in analysis?
        - » Valid only if no changes to protocol, and
        - » We would need to adjust for the stopping rule

95

Phase I Clinical Trials

- Initial safety / dose finding in humans
  - Goals:
    - Pharmacokinetics / pharmacodynamics
    - Incidence of major adverse effects
    - Decide whether it is ethical to continue testing in humans
  - Methods
    - Relatively small number of participants
    - Participants often not true target population
    - Sometimes dose escalation
    - Sometimes no comparison group

96

Phase IV Clinical Trials

- Therapeutic: Post-marketing surveillance
  - Goals:
    - Monitor for rare serious events
    - (Some "Phase IV" trials are of more interest for marketing than for science)
- Prevention: Effectiveness

97

Example

Series of RCT

Where am I going?

The investigation of new treatments, preventive strategies, and diagnostic procedures typically progresses through several phases.

This example illustrates decisions that might be made between Phase II and Phase III

98

Example: ROC HS/D Shock Trial

- Resuscitation Outcomes Consortium
  - 11 Geographic sites serving ~ 20 million
    - University based investigators
  - More than 250 EMS agencies
    - Over 35,000 EMS providers: EMTs and paramedics
- Conduct definitive clinical trials in the resuscitation of pre-hospital cardiac arrest and severe traumatic injury
  - Treat patients 20-50 minutes on average before delivering them to ED / hospital

ROC
RESUSCITATION OUTCOMES CONSORTIUM

99

Hypertonic Resuscitation in Shock

- Hypotheses: Use of hypertonic fluids (instead of normal saline) in patients with hypovolemic shock
  - Osmotic action to maintain fluid in vascular space
  - Anti-inflammatory effect to minimize reperfusion injury
- Randomized, double blind clinical trial
  - Hypotensive subjects following trauma receive 250 ml bolus of
    - 7.5% NaCl
    - 7.5% NaCl with dextran
    - Normal saline
  - All other treatments per standard medical care

100

21 CFR 50.24

- Exception to informed consent for research in an emergency setting
  - Unmet need
  - Study *effectiveness* of a therapy with some preliminary evidence of possible benefit
  - Consent impossible
  - Scientific question cannot be addressed in another setting
  - Patients in trial stand chance of benefit
  - Independent physicians attest to above
  - Community consultation / notification
  - As soon as possible notify subjects / next of kin of participation and right to withdraw

101

Background: Phase II Study

- HS/D vs Lactate Ringers in shock from blunt trauma
  - Primary endpoint: ARDS free survival at 28 days
- Group sequential design
  - Planned maximal sample size: 400 patients (200 / arm)
- Interim results after 200 patients
  - 28 day ARDS-free survival : 54% with HSD, 64% with LRS
  - DMC recommendation: Stop for futility
    - Trial results have excluded the hypothesized treatment effect
- Subgroup analysis
  - Suggestion of a benefit in the 20% needing massive transfusions
    - 28 day ARDS-free survival: 13% with HSD, 0% with LRS
  - (Results must be quite unpromising in the other subgroup

102

Bulger, et al., *Arch Surg 2008 **143**(2): 139 - 148.*

ROC Phase III Study

- HS/D vs HS vs NS in shock from trauma
  - Primary endpoint: All cause survival at 28 days
  - Hypotheses: 69.2 % with HS/D or HS vs 64.6% with NS
- Eligibility criteria modified to try to exclude patients that do not require transfusion
  - Phase II study:
    - SBP < 90 mmHg
  - Modification from exploratory analyses of Phase II data:
    - SBP < 70 mmHg or
    - 70 mmHg < SBP < 90 mmHg and HR > 108

103

Sample Size

- Fixed sample study:
  - Type I error 0.0125 due to multiple comparisons
  - 3,726 subjects regardless of observed treatment effect
  - Statistical significance if 4.1% improvement at end
- Group sequential monitoring:
  - No increase in maximal sample size
  - Therefore will have slight decrease in power depending on stopping boundary that is chosen

104

Sample Size: Group Sequential Study

- Group sequential rule for efficacy:
  - "O'Brien-Fleming" rule known for "early-conservativism"
  - Maximal sample size 3,726

	N Accrue	Efficacy Boundary		
		Z	Crude Diff	Est (95% CI; One-sided P)
First	621	6.000	0.272	0.263 (0.183, 0.329); P < 0.0001
Second	1,242	4.170	0.134	0.129 (0.070, 0.181); P < 0.0001
Third	1,863	3.350	0.088	0.082 (0.035, 0.129); P = 0.0004
Fourth	2,484	2.860	0.065	0.060 (0.019, 0.102); P = 0.0025
Fifth	3,105	2.540	0.052	0.048 (0.010, 0.085); P = 0.0070
Sixth	3,726	2.290	0.042	0.040 (0.005, 0.078); P = 0.0130

105

Sample Size: Group Sequential Study

- Tentative group sequential rule for noninferiority:
  - DoD interested in lesser volume of fluid in battlefield if equivalent
  - Ultimately rejected by DMC due to lack of benefit for subjects

	N Accrue	Futility Boundary		
		Z	Crude Diff	Est (95% CI; One-sided P)
First	621	-4.000	-0.181	-0.172 (-0.238, -0.092); P > 0.9999
Second	1,242	-2.800	-0.090	-0.084 (-0.137, -0.026); P = 0.9973
Third	1,863	-1.800	-0.047	-0.041 (-0.088, 0.006); P = 0.9581
Fourth	2,484	-1.200	-0.027	-0.022 (-0.064, 0.019); P = 0.8590
Fifth	3,105	-0.700	-0.014	-0.010 (-0.048, 0.028); P = 0.7090
Sixth	3,726	-0.290	-0.005	-0.003 (-0.041, 0.032); P = 0.5975

106

Sample Size: Group Sequential Study

- Group sequential rule for futility:
  - Based on rejecting the hypothesized treatment effect
  - Tradeoffs between average sample size and loss of power

	N Accrue	Futility Boundary		
		Z	Crude Diff	Est (95% CI; One-sided P)
First	621	-2.148	-0.097	-0.088 (-0.154 -0.008); P = 0.9837
Second	1,242	-0.605	-0.019	-0.011 (-0.066, 0.045); P = 0.6684
Third	1,863	0.372	0.010	0.017 (-0.031, 0.063); P = 0.2591
Fourth	2,484	1.120	0.025	0.030 (-0.011, 0.072); P = 0.0738
Fifth	3,105	1.740	0.035	0.038 (0.001, 0.078); P = 0.0209
Sixth	3,726	2.276	0.042	0.043 (0.005, 0.080); P = 0.0125

107

Comparison of Average Sample Size

- Average number of subjects treated according to the true effect (benefit or harm) of the treatment

True Benefit / Harm	Average Sample Size (Power)			
	Fixed Sample	Efficacy Only	Efficacy / Noninferiority	Efficacy / Futility
0.10	3,726 (.999)	1,968 (.999)	1,968 (.999)	1,940 (.998)
0.06	3,726 (.841)	2,930 (.832)	2,929 (.832)	2,754 (.817)
0.03	3,726 (.267)	3,578 (.259)	3,535 (.259)	2,729 (.252)
0.00	3,726 (.012)	3,720 (.012)	3,264 (.012)	1,995 (.012)
-0.03	3,726 (.000)	3,726 (.000)	2.374 (.000)	1,473 (.000)
-0.06	3,726 (.000)	3,726 (.000)	1,710 (.000)	1,181 (.000)

108

Benefit of Sequential Sampling

- Group sequential design can maintain type I error and power while greatly improving average sample size
  - To maintain power exactly, need slight increase in maximal N
- Improving average sample size increases number of beneficial treatments found by a consortium
- Advantage of group sequential over other adaptive strategies
  - Generally just as efficient
  - Better able to provide inference ("better understood" per FDA)

109

Why Not "Scientific Adaptation"

- From Phase II to Phase III we modified patient population to try to remove nontransfused subjects
  - Subjects with low blood pressure due to fainting?
  - Subjects who died before treatment could be administered?
- Why not do this in the middle of a trial?
  - *A priori:* Need to confirm and provide inference for indication
- In hindsight: Phase III still showed increased mortality in this subgroup that is identified post-randomization
  - Should we have modified treatment and/or eligibility?
  - More conservative approach in (at least) exception to informed consent argues for careful evaluation of confusing results

110

Final Comments

- In a large, expensive study, it is well worth our time to carefully examine the ways we can best protect
  - Patients on the study
  - Patients who might be on the study
  - Patients who will not be on the study, but will benefit from new knowledge
  - Sponsor's economic interests in cost of trial
  - Eventual benefit to health care
  - Eventual benefit to health care costs
- Adaptation to interim trial results introduces complications, but they can often be surmounted using methods that are currently well understood

111

Well Understood Methods

Group Sequential Designs

Where am I going?

Borrowing terminology from the draft CDER/CBER Guidance, we first review the extent to which the "well understood" group sequential designs can be viewed as adaptive designs.

I show that almost all of the motivation for adaptation of the maximal sample size is easily addressed in the group sequential framework.

112

Statistical Planning

- Satisfy collaborators as much as possible
  - Discriminate between relevant scientific hypotheses
    - Scientific and statistical credibility
  - Protect economic interests of sponsor
    - Efficient designs
    - Economically important estimates
  - Protect interests of patients on trial
    - Stop if unsafe or unethical
    - Stop when credible decision can be made
  - Promote rapid discovery of new beneficial treatments

113

Sample Size Calculation

- Traditional approach
  - Sample size to provide high power to "detect" a particular alternative
- Decision theoretic approach
  - Sample size to discriminate between hypotheses
    - "Discriminate" based on interval estimate
    - Standard for interval estimate: 95%
      - Equivalent to traditional approach with 97.5% power

114

Issues

- Summary measure
  - Mean, geometric mean, median, proportion, hazard…
- Structure of trial
  - One arm, two arms, k arms
  - Independent groups vs cross over
  - Cluster vs individual randomization
  - Randomization ratio
- Statistic
  - Parametric, semi-parametric, nonparametric
  - Adjustment for covariates

115

Refining Scientific Hypotheses

- Scientific hypotheses are typically refined into statistical hypotheses by identifying some parameter θ measuring difference in distribution of response
  - Difference/ratio of means
  - Ratio of geometric means
  - Difference/ratio of medians
  - Difference/ratio of proportions
  - Odds ratio
  - Hazard ratio

116

Inference

- Generalizations from sample to population
- Estimation
  - Point estimates
  - Interval estimates
- Decision analysis (testing)
  - Quantifying strength of evidence

117

Measures of Precision

- Estimators are less variable across studies
  - Standard errors are smaller
- Estimators typical of fewer hypotheses
  - Confidence intervals are narrower
- Able to statistically reject false hypotheses
  - Z statistic is higher under alternatives

118

Without Loss of Generality

- It is sufficient to consider a one sample test of a one-sided hypothesis
  - Generalization to other probability models is immediate
    - We will interpret our variability relative to average statistical info
  - Generalization to two sided hypothesis tests is straightforward
- Fixed sample one-sided tests
  - Test of a one-sided alternative ($\theta_+ > \theta_0$)
    - Upper Alternative: $H_+: \theta \geq \theta_+$ (superiority)
    - Null: $H_0: \theta \leq \theta_0$ (equivalence, inferiority)
  - Decisions based on some test statistic T:
    - Reject H_0 (for H_+) $\Longleftrightarrow$ $T \geq c$
    - Reject H_+ (for H_0) $\Longleftrightarrow$ $T \leq c$

119

Notation

Potential data : $Y_1, Y_2, Y_3, \ldots, Y_{N_j}$

Probability model : $Y_i \overset{iid}{\sim} (\theta, V)$

Interim estimates : $\hat{\theta}_{N_j} = \hat{\theta}(Y_1, \ldots, Y_{N_j})$

Without sequential sampling :

Approximate distn : $\hat{\theta}_j = \hat{\theta}_{N_j} \sim N(\theta, V / N_j)$

Indep increments : $Cov(\hat{\theta}_{N_j}, \hat{\theta}_{N_{j+1}}) = V / N_{j+1}$

Interim test statistics : $Z_j = Z_{N_j} = \dfrac{\hat{\theta}_j - \theta_0}{\sqrt{V / N_j}}$

120

Std Errors: Key to Precision

- Greater precision is achieved with smaller standard errors

Typically: $se(\hat{\theta}) = \sqrt{\frac{V}{n}}$

(V related to average "statistical information")

Width of CI: $2 \times (crit\ val) \times se(\hat{\theta})$

Test statistic: $Z = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$

121

Ex: One Sample Mean

$iid\ Y_i \sim (\mu, \sigma^2), i = 1, \ldots, n$

$\theta = \mu \qquad \hat{\theta} = \bar{Y}$

$V = \sigma^2 \qquad se(\hat{\theta}) = \sqrt{\frac{\sigma^2}{n}}$

122

Ex: Difference of Indep Means

$ind\ Y_{ij} \sim (\mu_i, \sigma_i^2), i = 1,2;\ j = 1, \ldots, n_i$

$n = n_1 + n_2; \quad r = n_1 / n_2$

$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$

$V = (r+1)[\sigma_1^2 / r + \sigma_2^2] \qquad se(\hat{\theta}) = \sqrt{\frac{V}{n}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

123

Ex: Diff of Indep Proportions

$ind\ Y_{ij} \sim B(1, p_i), i = 1,2;\ j = 1, \ldots, n_i$

$n = n_1 + n_2; \quad r = n_1 / n_2$

$\theta = p_1 - p_2 \qquad \hat{\theta} = \hat{p}_1 - \hat{p}_2 = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$

$\sigma_i^2 = p_i(1 - p_i)$

$V = (r+1)[\sigma_1^2 / r + \sigma_2^2] \qquad se(\hat{\theta}) = \sqrt{\frac{V}{n}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

124

Ex: Difference of Paired Means

$$Y_{ij} \sim \left(\mu_i, \sigma_i^2\right), i = 1,2;\ j = 1,\ldots,n$$
$$corr\left(Y_{1j}, Y_{2j}\right) = \rho; \quad corr\left(Y_{ij}, Y_{mk}\right) = 0 \text{ if } j \neq k$$
$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$$

$$V = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}}$$

125

Ex: Mean of Clustered Data

$$Y_{ij} \sim \left(\mu, \sigma^2\right), i = 1,\ldots,n;\ j = 1,\ldots,m$$
$$corr\left(Y_{ij}, Y_{ik}\right) = \rho \text{ if } j \neq k; \quad corr\left(Y_{ij}, Y_{mk}\right) = 0 \text{ if } i \neq m$$
$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$$

$$V = \sigma^2\left(\frac{1 + (m-1)\rho}{m}\right) \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}}$$

126

Ex: Independent Odds Ratios

$$ind\ Y_{ij} \sim B\left(1, p_i\right), i = 1,2;\ j = 1,\ldots,n_i$$
$$n = n_1 + n_2; \quad r = n_1 / n_2$$
$$\theta = \log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right) \qquad \hat{\theta} = \log\left(\frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}\right)$$
$$\sigma_i^2 = \frac{1}{p_1(1-p_1)} = \frac{1}{p_1 q_1}$$
$$V = (r+1)\left[\sigma_1^2/r + \sigma_2^2\right] \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}} = \sqrt{\frac{1}{n_1 p_1 q_1} + \frac{1}{n_2 p_2 q_2}}$$

127

Ex: Hazard Ratios

$$ind \text{ censored time to event } \left(T_{ij}, \delta_{ij}\right),$$
$$i = 1,2;\ j = 1,\ldots,n_i; n = n_1 + n_2; \quad r = n_1 / n_2$$
$$\theta = \log(HR) \qquad \hat{\theta} = \hat{\beta} \text{ from PH regression}$$

$$V = \frac{(1+r)(1/r+1)}{\Pr\left[\delta_{ij} = 1\right]} \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}} = \sqrt{\frac{(1+r)(1/r+1)}{d}}$$

128

Ex: Linear Regression

$$ind\ Y_i \mid X_i \sim \left(\beta_0 + \beta_1 \times X_i, \sigma^2_{Y|X}\right), i = 1, \ldots, n$$

$$\theta = \beta_1 \qquad \hat{\theta} = \hat{\beta}_1 \text{ from LS regression}$$

$$V = \frac{\sigma^2_{Y|X}}{Var(X)} \qquad se(\hat{\theta}) = \sqrt{\frac{\sigma^2_{Y|X}}{nVar(X)}}$$

129

Statistics to Address Variability

- At the end of the study we perform frequentist and/or Bayesian data analysis to assess the credibility of clinical trial results
  - Estimate of the treatment effect
    - Single best estimate
    - Precision of estimates
  - Decision for or against hypotheses
    - Binary decision
    - Quantification of strength of evidence

130

Criteria for Precision

- Standard error
- Width of confidence interval
- Statistical power: Probability of rejecting the null hypothesis
  - Select level of significance
    - Standard: One-sided 0.025; two-sided 0.05
    - Pivotal: One-sided 0.005; two-sided 0.01
  - Select "design alternative"
    - Minimal clinically important difference
      - To detect versus declaring significant
      - May consider what is feasible
    - Minimal plausible difference
  - Select desired power
    - High power for a decision theoretic approach

131

Sample Size Determination

- Based on sampling plan, statistical analysis plan, and estimates of variability, compute
  - Sample size that discriminates hypotheses with desired power,

    OR
  - Hypothesis that is discriminated from null with desired power when sample size is as specified, or

    OR
  - Power to detect the specific alternative when sample size is as specified

132

Sample Size Computation

Standardized level α test (n = 1): $\delta_{\alpha\beta}$ detected with power β

Level of significance α when $\theta = \theta_0$

Design alternative $\theta = \theta_1$

Variability V within 1 sampling unit

Required sampling units : $$n = \frac{(\delta_{\alpha\beta})^2 V}{(\theta_1 - \theta_0)^2}$$

(Fixed sample test : $\delta_{\alpha\beta} = z_{1-\alpha/2} + z_{\beta}$)

133

When Sample Size Constrained

- Often (usually?) logistical constraints impose a maximal sample size
  - Compute power to detect specified alternative

Find β such that $$\delta_{\alpha\beta} = \sqrt{\frac{n}{V}}(\theta_1 - \theta_0)$$

  - Compute alternative detected with high power

$$\theta_1 = \theta_0 + \delta_{\alpha\beta}\sqrt{\frac{V}{n}}$$

134

Increasing Precision

- Options
  - Increase sample size
  - Decrease V
    - Improve reliability of measurements
    - Alter study design (e.g., cross-over)
    - Alter eligibility (decrease heterogeneity)
    - Alter clinical endpoint
  - (Decrease confidence level)

135

Additional Constraints

- Safety analyses
  - Often there is a minimal number needed to treat in order to have enough data to rule out unacceptably high rates of extremely serious adverse events
    - "3 over n rule" as confidence bound when no such events observed
- Subgroup analyses
  - May need sufficient data to examine effects in important subgroups

136

Without Loss of Generality

- Our ultimate interest is in comparing
  - Fixed sample tests
  - Group sequential tests
  - Other adaptive strategies
- We will thus further restrict attention to a one-sample setting in which
  - $V = 1$
  - Test of a one-sided alternative ($\theta_+ > \theta_0$)
    - Upper Alternative: $H_+: \theta \geq \theta_+ = 3.92$ (superiority)
    - Null: $H_0: \theta \leq \theta_0 = 0$ (equivalence, inferiority)

137

Fixed Sample Test

- Sample size N = 1 provides
  - Type 1 error of 0.025
  - Power of 0.975 to detect the alternative of 3.92
  - At the final analysis, an observed estimate (or Z statistic) of 1.96 will be statistically significant
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	1.00
1.96	0.500	1.00
2.80	0.800	1.00
3.24	0.900	1.00
3.92	0.975	1.00

138

Group Sequential Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - Can be randomly determined if independent of effect
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
  - Often chosen according to some boundary shape function
    - O'Brien-Fleming, Pocock, Triangular, …
- Compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue

139

Stopping Boundary Scales

- Boundary scales (1:1 transformations among these)
  - Z statistic
  - P value
    - Fixed sample (so wrong)
    - Computed under sequential sampling rule (so correct)
  - Error spending function
  - Estimates
    - MLE (biased due to stopping rule)
    - Adjusted for stopping rule
  - Conditional power
    - Computed under design alternative
    - Computed under current MLE
  - Predictive power
    - Computed under flat prior (possibly improper)

140

Exploring Group Sequential Designs

- Candidate designs
  - J = 2 equal spaced analyses; O'Brien-Fleming efficacy boundary
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 2 equal spaced analyses; OBF efficacy, futility boundaries
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 2 equal spaced analyses; OBF efficacy, more efficient futility
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 4 equal spaced analyses; OBF efficacy, more efficient futlility
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 2 optimally spaced analyses; optimal symmetric boundaries
    - Maintain power under alternative (inflate N_J, but optimize ASN)

141

Exploring Group Sequential Designs

- Examining operating characteristics
  - Stopping boundaries
    - Z scale
    - Conditional power under hypothesized effects
    - Conditional power under current MLE
    - Predictive power under flat prior
  - Estimates and inference
    - MLE (Bias adjusted estimates suppressed for space)
    - 95% CI properly adjusted for stopping rule
    - P value properly adjusted for stopping rule
  - Power at specified alternatives
  - Sample size distribution (as function of true effect)
    - Maximal sample size
    - Average sample size

142

O'Brien-Fleming Efficacy: J = 2

- Introduce two evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.5	--	--	--	--	2.796	0.500	0.997	0.976
1.0	1.977	--	--	--	1.977	--	--	--

143

O'Brien-Fleming Efficacy: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.5	--	--	--	3.955	(1.16, 5.72)	0.003
1.0	1.977	(0.00, 3.93)	0.025	1.977	(0.00, 3.93)	0.025

144

O'Brien-Fleming Efficacy: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.999
1.96	0.496	0.960
2.80	0.797	0.896
3.24	0.898	0.847
3.92	0.974	0.755

145

O'Brien-Fleming Efficacy: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.50	--	--	--	3.943	(1.16, 5.70)	0.003
1.01	1.977	(0.00, 3.92)	0.025	1.977	(0.00, 3.92)	0.025

146

O'Brien-Fleming Efficacy: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	1.005
1.96	0.499	0.966
2.80	0.799	0.901
3.24	0.900	0.851
3.92	0.975	0.758

147

Take Home Messages 1

- Introduction of a very conservative efficacy boundary
  - Minimal effect on power even if do not increase max N
  - Minimal increase in max N needed to maintain power
- Ease and importance of evaluating a stopping rule
  - Even before we start the study, we can consider
    - Thresholds for early stopping in terms of estimated effects
    - Inference corresponding to stopping points
    - Conditional and predictive power under various hypotheses
  - We can judge a stopping rule by comparing it to a fixed sample test and look at the tradeoffs between
    - Increase in maximal sample size
    - Decrease in average sample size
    - Changes in unconditional power

148

O'Brien-Fleming Symmetric: J = 2

- Introduce two evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.5	0.000	0.500	0.003	0.024	2.796	0.500	0.997	0.976
1.0	1.977	--	--	--	1.977	--	--	--

149

O'Brien-Fleming Symmetric: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.5	0.000	(-1.76, 2.80)	0.375	3.945	(1.15, 5.71)	0.003
1.0	1.973	(0.00, 3.94)	0.025	1.973	(0.00, 3.94)	0.025

150

O'Brien-Fleming Symmetric: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.749
1.96	0.495	0.919
2.80	0.795	0.883
3.24	0.897	0.840
3.92	0.974	0.752

151

O'Brien-Fleming Symmetric: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.51	0.00	(-1.75, 2.78)	0.375	3.920	(1.14, 5.67)	0.003
1.01	1.960	(0.00, 3.92)	0.025	1.960	(0.00, 3.92)	0.025

152

O'Brien-Fleming Symmetric: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.758
1.96	0.500	0.930
2.80	0.800	0.893
3.24	0.900	0.848
3.92	0.975	0.758

153

Take Home Messages 2

- Introduction of a very conservative futility boundary
  - Again, minimal effects on power and/or max N
  - Dramatic improvement in ASN under the null
  - Conditional and predictive power thresholds are surprising
    - CP_{alt} = 0.50 for the extremely conservative OBF boundary
      - But the CI has already eliminated 3.92 with high confidence
    - CP_{est} = 0.003 and PP_{flat} = 0.024 are both very low thresholds

154

O'Brien-Fleming & Futility: J = 2

- Introduce two evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.5	0.331	0.644	0.017	0.068	2.776	0.500	0.997	0.975
1.0	1.963	--	--	--	1.963	--	--	--

155

O'Brien-Fleming & Futility: J = 2, N = 1

- Introduce four evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.5	0.468	(-1.30, 3.27)	0.228	3.925	(1.13, 5.69)	0.003
1.0	1.963	(0.00, 3.98)	0.025	1.963	(0.00, 3.98)	0.025

156

O'Brien-Fleming & Futility: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.684
1.96	0.492	0.886
2.80	0.791	0.869
3.24	0.893	0.830
3.92	0.972	0.747

157

O'Brien-Fleming & Futility: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.52	0.461	(-1.28, 3.22)	0.228	3.867	(1.11, 5.61)	0.003
1.03	1.934	(0.00, 3.92)	0.025	1.934	(0.00, 3.92)	0.025

158

O'Brien-Fleming & Futility: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.705
1.96	0.504	0.914
2.80	0.803	0.892
3.24	0.901	0.850
3.92	0.975	0.762

159

Take Home Messages 3

- More aggressive futility boundary better addresses ethical issues associated with ineffective drugs
  - I often find that sponsors are willing to accept this futility bound without increasing the sample size
  - But the minimal increase in maximal sample size would seem more appropriate to me

160

O'Brien-Fleming & Futility: J = 4

- Introduce four evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.25	-1.108	0.719	0.000	0.008	3.976	0.500	0.999	0.999
0.50	0.321	0.648	0.015	0.063	2.811	0.500	0.997	0.977
0.75	1.258	0.592	0.142	0.177	2.295	0.500	0.907	0.874
1.00	1.988	--	--	--	1.988	--	--	--

161

O'Brien-Fleming & Futility: J = 4, N = 1

- Introduce four evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.25	-2.216	(-4.71, 1.74)	0.846	7.951	(4.00, 10.5)	0.000
0.50	0.454	(-1.60, 3.31)	0.263	3.976	(1.14, 6.04)	0.003
0.75	1.452	(-0.36, 3.85)	0.053	2.650	(0.30, 4.48)	0.013
1.00	1.988	(0.00, 4.06)	0.025	1.988	(0.00, 4.06)	0.025

162

O'Brien-Fleming & Futility: J = 4, N = 1

- Introduce four evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.580
1.96	0.478	0.783
2.80	0.776	0.761
3.24	0.882	0.723
3.92	0.966	0.650

163

O'Brien-Fleming & Futility: J = 4, Power

- Introduce four evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.27	-2.141	(-4.55, 1.68)	0.846	7.682	(3.86, 10.1)	0.000
0.54	0.439	(-1.55, 3.20)	0.263	3.841	(1.10, 5.84)	0.003
0.80	1.403	(-0.34, 3.72)	0.053	2.561	(0.29, 4.33)	0.013
1.07	1.920	(0.00, 3.92)	0.025	1.920	(0.00, 3.92)	0.025

164

O'Brien-Fleming & Futility: J = 4, Power

- Introduce four evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.622
1.96	0.504	0.840
2.80	0.803	0.808
3.24	0.902	0.762
3.92	0.975	0.680

165

Take Home Messages 4

- Effect of adding more analyses
  - Greater loss of power if maximal sample size not increased
  - Greater increase in maximal sample size if power maintained
  - But, improvement in average efficiency
- Can also use this example for guidance in how to judge thresholds for conditional and predictive power
  - The same threshold should not be used at all analyses
  - It is not, however, clear what threshold should be used
    - I look at tradeoffs between average efficiency and power
  - We can look at optimal (on average) designs for more guidance

166

Efficient: J = 2

- Introduce two optimally spaced analyses to minimize ASN
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.43	0.573	0.818	0.049	0.141	2.776	0.182	0.951	0.859
1.00	2.129	--	--	--	2.129	--	--	--

167

Efficient: J = 2, Power

- Introduce two optimally spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.50	0.808	(-0.82, 3.58)	0.129	3.112	(0.34, 4.74)	0.014
1.18	1.960	(0.00, 3.92)	0.025	1.960	(0.00, 3.92)	0.025

168

Efficient: J = 2, Power

- Introduce two optimally spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.685
1.96	0.500	0.900
2.80	0.805	0.847
3.24	0.904	0.788
3.92	0.975	0.685

169

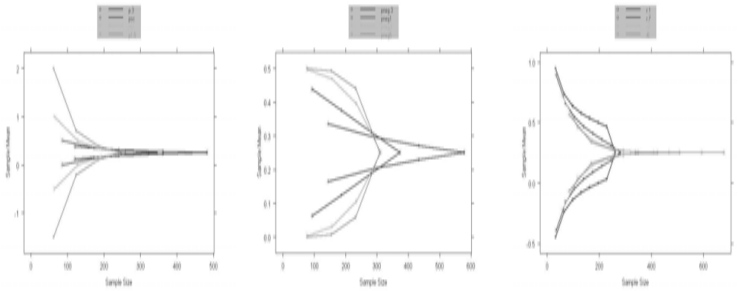
Take Home Messages 5

- Optimal spacing of analyses not quite equal information
- Optimal early conservatism close to a Pocock design
  - In unified family, OBF has P=1, Pocok has P= 0.5
  - Optimal P= .54
- With two analyses, increase maximal N by 18% over fixed sample
  - ASN decreases by about one third
- Again, the thresholds to use for conditional or predictive power are not at all clear
- Search for best designs should include many candidates
  - Examine many operating characteristics

170

Spectrum of Boundary Shapes

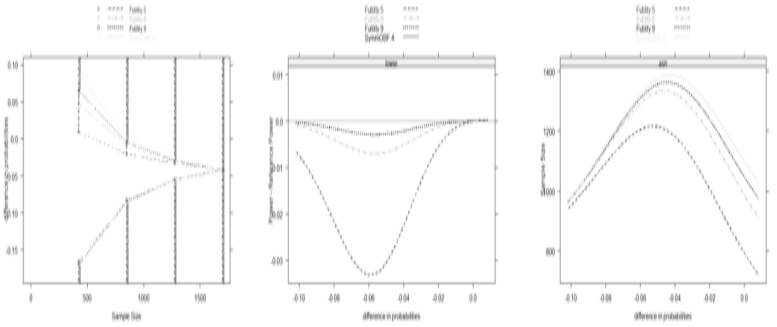
- All of the rules depicted have the same type I error and power to detect the design alternative



Efficiency / Unconditional Power

- Tradeoffs between early stopping and loss of power

Boundaries Loss of Power Avg Sample Size



Delayed Measurement of Outcome

- Longitudinal studies
  - Measurement might be 6 months – 2 years after randomization
  - Interim analyses on variable lengths of follow-up
    - Use of partial data can improve efficiency (Kittelson, et al.)
- Time to event studies
  - Statistical information proportional to number of events
  - Calendar time requirements depend on number accrued and length of follow-up
- In either case: Interim analyses may occur after accrual completed
    - Group ethics of identifying beneficial treatments faster
    - Savings in calendar time costs, rather than per patient costs

173

Response Adaptive Modifications of an RCT Design

Overview of General Methods

Where am I going?

Various authors have advocated the use of unblinded interim estimates of the treatment effect to modify a broad spectrum of RCT design parameters.

In this section of this course, I review the most commonly espoused statistical methods that would be used to maintain an experimentwise type I error.

174

Blinded Adaptive Sampling

- Modify sample size to account for estimated information (variance or baseline rates)
- No effect on type I error IF
  - Estimated information independent of estimate of treatment effect
    - Proportional hazards,
    - Normal data, and/or
    - Carefully phrased alternatives
  - And willing to use conditional inference
    - Carefully phrased alternatives

175

Estimation of Statistical Information

- If maximal sample size is maintained, the study discriminates between null hypothesis and an alternative measured in units of statistical information

$$n = \frac{\delta_1^2 V}{(\Delta_1 - \Delta_0)^2} \qquad n = \frac{\delta_1^2}{\left(\frac{(\Delta_1 - \Delta_0)^2}{V}\right)}$$

176

Estimation of Statistical Information

- If statistical power is maintained, the study sample size is measured in units of statistical information

$$n = \frac{\delta_1{}^2 V}{(\Delta_1 - \Delta_0)^2} \qquad \frac{n}{V} = \frac{\delta_1{}^2}{(\Delta_1 - \Delta_0)^2}$$

177

Adaptation to Gain Efficiency?

- Consider adaptation merely to repower study
  - "We observed a result that was not as good as we had anticipated"
- All GST are within family of adaptive designs
  - Don't we have to be at least as efficient?
- Issues
  - Unspecified adaptations
  - Comparing apples to apples

178

Two Stage Design

- Proschan & Hunsberger consider worst case
  - At first stage, choose sample size of second stage
    - $N_2 = N_2(Z_1)$ to maximize type I error
  - At second stage, reject if $Z_2 > a_2$

179

Proschan & Hunsberger

- Worst case type I error of two stage design

$$\alpha_{worst} = 1 - \Phi\left(a_2^{(Z)}\right) + \frac{\exp\left(-\left(a_2^{(Z)}\right)^2 / 2\right)}{4},$$

- Can be more than two times the nominal
  - $a_2 = 1.96$ gives type I error of 0.0616
  - (Compare to Bonferroni results)

180

Flexible Adaptive Designs

- Proschan and Hunsberger describe adaptations to maintain experimentwise type I error and increase conditional power
  - Must prespecify a conditional error function

$$\int_{-\infty}^{\infty} A(z)\,\phi(z)\,dz = \alpha.$$

  - Often choose function from some specified test

$$A(z) = Pr_{\delta=0}(Z_2 \geq \Phi^{-1}(1-\alpha)\,|\,\tilde{Z}_1 = z, \tilde{n}_2 = n_2 - n_1),$$

  - Find critical value to maintain type I error

$$Pr_{\delta=0}\,(Z_2^* \geq c(\tilde{n}_2^*, \tilde{z}_1)\,|\,\tilde{n}_2^*(\tilde{z}_1)) = A(\tilde{z}_1).$$

181

Other Approaches

- Self-designing Trial (Fisher, 1998)
  - Combine arbitrary test statistics from sequential groups
  - Prespecify weighting of groups "just in time"
    - Specified at immediately preceding analysis
  - Fisher's test statistic is N(0,1) under the null hypothesis of no treatment difference on any of the endpoints tested
- Combining P values (Bauer & Kohne, 1994)
  - Based on R.A. Fisher's method

182

Incremental Statistics

- Statistic at the j-th analysis a weighted average of data accrued between analyses

$$N_k^* = N_k - N_{k-1}$$

Statistics computed on kth increment : $\hat{\theta}_k^* \quad Z_k^* \quad P_k^*$

$$\hat{\theta}_j = \frac{\sum_{k=1}^{j} N_k^* \hat{\theta}_k^*}{N_j} \qquad Z_j = \frac{\sum_{k=1}^{j} \sqrt{N_k^*}\, Z_k^*}{\sqrt{N_j}}.$$

183

Conditional Distribution

$$\hat{\theta}_j^* \mid N_j^* \sim N\left(\theta, \frac{V}{N_j^*}\right)$$

$$Z_j^* \mid N_j^* \sim N\left(\frac{\theta - \theta_0}{\sqrt{V / N_j^*}}, 1\right)$$

$$H_0$$

$$P_j^* \mid N_j^* \sim U(0,1).$$

184

Protecting Type I Error

- LD Fisher's variance spending method
  - Arbitrary hypotheses H_{0j}:$\theta_j = \theta_{0j}$
  - Incremental test statistics Z_j^*
  - Allow arbitrary weights W_j specified at stage *j-1*

$$Z_j = \frac{\sum_{k=1}^{J} \sqrt{W_k}\ Z_k^*}{\sqrt{\sum_{k=1}^{J} W_j}}$$

- RA Fisher's combination of P values (Bauer & Köhne)

$$P_j = \prod_{k=1}^{j} P_j^*$$

185

Unconditional Distribution

- Under the null
  - SDCT: Standard normal
  - Bauer & Kohne: Sum of exponentials
- Under the alternative
  - Unknown unless prespecified adaptations

$$\Pr\left(Z_j^* \le z\right) = \sum_{n=0}^{\infty} \Pr\left(Z_j^* \le z \mid N_j^*\right) \Pr\left(N_j^* = n\right).$$

186

Sufficiency Principle

- It is easily shown that a minimal sufficient statistic is (Z, N) at stopping
- All methods advocated for adaptive designs are thus not based on sufficient statistics

187

Response Adaptive Modifications of an RCT Design

Modification of the Maximal Sample Size

Where am I going?

The majority of the statistical literature on adaptive clinical trial design is directed toward using interim estimates of the treatment effect to re-power a study.

In this section of the course I review the basis for those methods and demonstrate the settings in which such adaptive designs can improve substantially on more traditional methods.

188

Blinded Adaptive Sampling

- Modify sample size to account for estimated information (variance or baseline rates)
- No effect on type I error IF
  - Estimated information independent of estimate of treatment effect
    - Proportional hazards,
    - Normal data, and/or
    - Carefully phrased alternatives
  - And willing to use conditional inference
    - Carefully phrased alternatives

189

Estimation of Statistical Information

- If maximal sample size is maintained, the study discriminates between null hypothesis and an alternative measured in units of statistical information

$$n = \frac{\delta_1^{\,2} V}{(\Delta_1 - \Delta_0)^2} \qquad n = \frac{\delta_1^{\,2}}{\left(\frac{(\Delta_1 - \Delta_0)^2}{V}\right)}$$

190

Estimation of Statistical Information

- If statistical power is maintained, the study sample size is measured in units of statistical information

$$n = \frac{\delta_1^{\,2} V}{(\Delta_1 - \Delta_0)^2} \qquad \frac{n}{V} = \frac{\delta_1^{\,2}}{(\Delta_1 - \Delta_0)^2}$$

191

Adaptive Sample Size Determination

- Design stage:
  - Choose an interim monitoring plan
  - Choose an adaptive rule for maximal sample size
- Conduct stage:
  - Recruit subjects, gather data in groups
  - After each group, analyze for DMC
    - DMC recommends termination or continuation
  - After penultimate group, determine final N
- Analysis stage:
  - When study stops, analyze and report

192

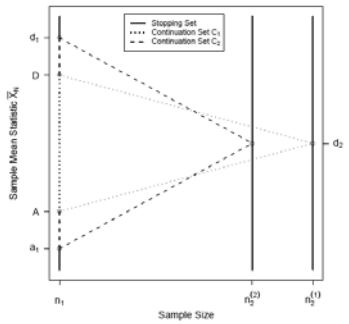
Adaptive Sample Size Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - $N_1 \ldots N_{J-1}$ can be randomly determined indep of effect
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- At $N_1 \ldots N_{J-1}$ compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue
    - At N_{J-1} determine N_J according to value of T_j

193

Operating Characteristics

- Given an adaptive rule that modifies the final sample size at the penultimate stage, this can be viewed as adaptively switching between two group sequential tests
  - Standard group sequential software can be used



194

Apples with Apples

- Can adapting beat a GST with the same number of analyses?
  - Fixed sample design: N=1
  - Most efficient symmetric GST with two analyses
    - N = 0.5, 1.18
    - ASN = 0.6854
  - Most efficient adaptive design with two possible N
    - N = 0.5 and either 1.06 or 1.24
    - ASN = 0.6831 (0.34% more efficient)
  - "Most efficient" adaptive design with four possible N
    - N = 0.5 and either 1.01, 1.10, 1.17, or 1.31
    - ASN = 0.6825 (0.42% more efficient)

Table 1: Average and Maximal Sample Sizes of Adaptive Designs in Setting 1

	Number of Continuation Regions							
	1	2	3	4	5	6	7	8
ASN	0.6854	0.6831	0.6828	0.6825	0.6824	0.6824	0.6824	0.6824
% Reduction	Ref	0.34%	0.38%	0.42%	0.43%	0.43%	0.44%	0.44%
Maximal N	1.18	1.24	1.24	1.26	1.26	1.26	1.26	1.28

195

Apples with Apples (continued)

- GST with more analyses?
  - Fixed sample design: N=1
  - Most efficient symmetric GST with two analyses
    - N = 0.5, 1.18
    - ASN = 0.6854
  - GST with same three analyses
    - N = 0.5,1.06 and 1.24
    - ASN = 0.6666 (2.80% more efficient)
  - GST with same five analyses
    - N = 0.5, 1.01, 1.10, 1.17, or 1.31
    - ASN = 0.6576 (4.20% more efficient)

196

Prespecified Modification Rules

- Adaptive sampling plans exact a price in statistical efficiency
- Tsiatis & Mehta (2002); Jennison & Turnbull (2003)
  - A classic prespecified group sequential stopping rule can be found that is more efficient than a given adaptive design
- Shi (2003, unpublished thesis)
  - Fisher's test statistic in the self-designing trial provides markedly less precise inference than that based on the MLE
    - To compute the sampling distribution of the latter, the sampling plan must be known

197

Comments re Conditional Power

- Many authors propose adaptations based on conditional or predictive power
- Neither conditional power nor predictive power have good foundational motivation
  - Frequentists should use Neyman-Pearson paradigm and consider optimal unconditional power across alternatives
    - And conditional/predictive power is not a good indicator in loss of unconditional power
  - Bayesians should use posterior distributions for decisions
- Difficulty understanding conditional / predictive power scales can lead to bad choices for designs

198

Example

Adaptive Increase in Sample Size when Interim Results are Promising: A Practical Guide with Examples

Cyrus R. Mehta[1,2], Stuart J. Pocock[3]

[1]*Cytel Corporation,* [2]*Harvard School of Public Health,* [3]*London School of Hygiene and Tropical Medicine*

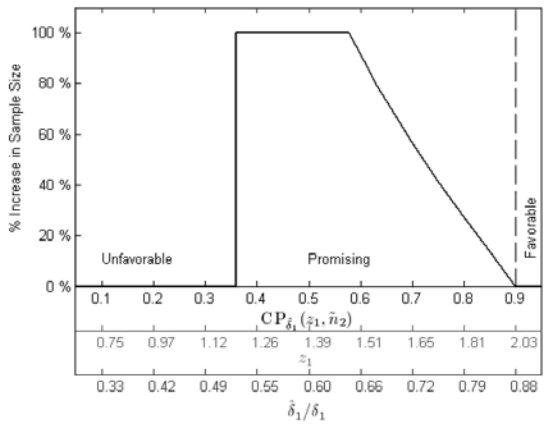
SUMMARY

This paper discusses the benefits and limitations of adaptive sample size re-estimation for phase 3 confirmatory clinical trials. Comparisons are made with more traditional fixed sample and group sequential designs. It is seen that the real benefit of the adaptive approach arises through the ability to invest sample size resources into the trial in stages. The trial starts with a small up-front sample size commitment. Additional sample size resources are committed to the trial only if promising results are obtained at an interim analysis. This strategy is shown through examples of actual trials, one in neurology and one in cardiology, to be more advantageous than the fixed sample or group sequential approaches in certain settings. A major factor that has generated controversy and inhibited more widespread use of these methods has been their reliance on non-standard tests and p-values for preserving the type-1 error. If, however, the sample size is only increased when interim results are promising, one can dispense with these non-standard methods of inference. Therefore, in the spirit of making adaptive increases in trial size more widely appealing and readily implementable we here define those promising circumstances in which a conventional final inference can be performed while preserving the overall type-1 error. Methodological, regulatory and operational issues are examined.

- http://www.cytel.com/pdfs/Mehta_Pocock_PromisingZone_StatsinMed_9.11.10.pdf

199

Example Modification Plan



200

Alternative Approaches

Table 1: Comparison of RCT Designs for Example 1

Design	Hypothesized Treatment Effect $\delta = 0$	$\delta = 1.5$	$\delta = 1.6$	$\delta = 1.7$	$\delta = 1.8$	$\delta = 1.9$	$\delta = 2.0$
				Power			
Fxd442	2.5%	55.6%	61.1%	66.3%	71.3%	75.9%	80.0%
Fxd690	2.5%	74.8%	80.0%	84.5%	88.3%	91.4%	93.9%
GST694	2.5%	74.8%	80.0%	84.6%	88.4%	91.4%	93.9%
Adapt	2.5%	60.4%	65.8%	70.8%	75.4%	79.6%	83.4%
Fxd492	2.5%	60.2%	65.8%	71.0%	75.9%	80.2%	84.1%
Fut492	2.5%	59.8%	65.4%	70.6%	75.4%	79.8%	83.7%
OBF492	2.5%	59.6%	65.2%	70.4%	75.3%	79.6%	83.5%
				Expected Number Accrued			
Fxd442	442	442	442	442	442	442	442
Fxd690	690	690	690	690	690	690	690
GST694	694	681	678	675	671	667	662
Adapt	464	496	495	494	492	490	488
Fxd492	492	492	492	492	492	492	492
Fut492	468	488	489	490	490	490	491
OBF492	467	485	485	485	485	484	484

201

Alternative Approaches

Table 1: Comparison of RCT Designs for Example 1

Design	Hypothesized Treatment Effect $\delta = 0$	$\delta = 1.5$	$\delta = 1.6$	$\delta = 1.7$	$\delta = 1.8$	$\delta = 1.9$	$\delta = 2.0$
				Expected Number Completed			
Fxd442	442	442	442	442	442	442	442
Fxd690	690	690	690	690	690	690	690
GST694	693	668	663	657	649	641	632
Adapt	464	496	495	494	492	490	488
Fxd492	492	492	492	492	492	492	492
Fut492	353	472	475	478	481	483	485
OBF492	352	455	455	454	452	449	445
				Expected Calendar Time (months)			
Fxd442	18.8	18.8	18.8	18.8	18.8	18.8	18.8
Fxd690	25.9	25.9	25.9	25.9	25.9	25.9	25.9
GST694	26.0	25.3	25.1	24.9	24.7	24.5	24.2
Adapt	19.4	20.3	20.3	20.3	20.2	20.1	20.1
Fxd492	20.2	20.2	20.2	20.2	20.2	20.2	20.2
Fut492	16.2	19.6	19.7	19.8	19.9	19.9	20.0
OBF492	16.1	19.1	19.1	19.1	19.0	19.0	18.8

202

Alternative Approaches

- The authors plan for adaptation could increase sample size by 100%
- Using their adaptive plan, the probability of continuing until a 25% increase in maximal sample size
  - .064 under null hypothesis
  - .162 if treatment effect is new target of 1.6
  - .142 if treatment effect is old target of 2.0
- By way of contrast
  - A fixed sample test with 11% increase in sample size has same power
  - A group sequential test with 11% increase in maximal sample size has same power and better ASN

203

Apparent Problem

- The authors chose extremely inefficient thresholds for conditional power
  - Adaptation region $0.365 < CP_{est} < 0.8$
  - From optimal test, $0.049 < CP_{est} < 0.8$ is optimal
- Even experienced clinical trialists get it wrong
  - The authors also seemed somewhat chagrined that efforts to boost conditional power did not result in substantial gains in unconditional power

204

The Cost of Planning Not to Plan

- In order to provide frequentist estimation, we must know the rule used to modify the clinical trial
- Hypothesis testing of a null is possible with fully adaptive trials
  - Statistics: type I error is controlled
  - Game theory: chance of "winning" with completely ineffective therapy is controlled
  - Science:
    - At best: ability to discriminate clinically relevant hypothesis may be impaired
    - At worst: uncertainty as to what the treatment has effect on
- With prespecified adaptation, we know how to do inference, but standard orderings may lead to surprising results

205

Response Adaptive Modifications of an RCT Design

Modification of the Randomization Ratio

Where am I going?

Some authors have proposed adaptive clinical trial designs in which the proportion of subjects assigned to each treatment is modified according to interim estimates of the treatment effect.

In this section of the course I illustrate some of the controversies that these approaches have caused in the past.

206

Ethics

- Clinical trials are experiments in human volunteers
- Individual ethics
  - Patients on trial: Avoid continued administration of inferior treatment
  - Patients not yet on trial: Avoid starting inferior treatment
- Group ethics
  - Facilitate rapid adoption of new beneficial treatments
  - Avoid prolonging study of ineffective treatments

207

Solutions

- Most commonly used
  - Sequential sampling
    - Interim analyses of data
    - Terminate trials when credible decisions can be made
- Also proposed
  - Response adaptive randomization
    - Change randomization probabilities as evidence accumulates that one treatment might be best
    - "Play the winner"

208

Play the Winner: Urn Model

- Begin with *k* white balls and *k* black balls in an urn
- Upon accrual of a patient draw a ball from urn
  - White ➔ control; black ➔ treatment
- Observe outcome
  - If outcome is good, return *m+1* balls of same color as withdrawn
  - If outcome is bad, return 1 ball of same color as withdrawn and *m* balls of opposite color

209

Bayesian Methods

- An explicit Bayesian approach dynamic randomization could base the randomization ration on the current posterior probability that one treatment is superior
  - Ultimately, that posterior probability is based on the number of good outcomes on each treatment
- Advantage of using Bayesian posterior probability
  - Can easily handle continuous outcomes
  - Can easily handle continuous randomization probabilities

210

Analytic Issues

- Treatment of successive patients is not independent of previous patients treatment and results
  - Possible bias in accrual of future patients
  - Analysis must stratify on time of accrual
- Conditionally biased estimates of treatment effect in arm with lower sample sizes
  - Bad early results tend to preclude regression to mean
- Randomization hypothesis can lead to quite unconvincing results

211

Example: ECMO Study

- Randomized clinical trial of extracorporeal membrane oxygenation in newborns
  - Randomized PTW design with k=1
- Data:
  - First patient on ECMO survived
  - Next patient on control died
  - Next 9 patients on ECMO survived
- Inference (Begg, 1990)
  - P value of 0.001, 0.051, 0.083, 0.28, 0.62?

212

Comments

- This experience has tempered enthusiasm for randomized PTW
  - Interestingly, follow-up studies had 67% survival on conventional therapy
- I believe there can be times that this will work, but
  - There needs to be a clear dilemma re individual ethics
  - There will tend to be decreased group ethics
  - It takes a lot of planning in order to obtain results that will be sufficiently credible
    - Just assuming your conclusion will not cut it

213

Response Adaptive Modifications of an RCT Design

Modification of the Eligibility Criteria

Where am I going?

It is rare that every treatment would work equally well in all subgroups.

Investigators might want to use interim estimates of treatment effect within subgroups to enrich the sampling of the more promising subgroups and drop the least promising subgroups.

I discuss the settings in which such "adaptive enrichment" might lead to improved efficiency of the process of adopting new treatments.

214

Phase II vs Phase III

- Phase II studies:
  - Finding best
    - Treatment
    - Dose
    - Population
    - Measure of outcome
  - Adaptation has a very natural role
- Phase III studies: Confirmatory
  - Adaptation is more problematic

215

Example 1: History

- Industry sponsored trials of new biologic agent in the treatment of pneumonia
  - Investigation proceeded through Phase I, Phase II testing
  - Preliminary evidence of efficacy supported progressing to Phase III trial
  - In Phase III trial, some suggestion that treatment is most beneficial in subgroup of patients
    - Unclear which dose was most promising

216

Example 1: History

- Manufacturing process limitations
  - Unless a marked benefit of higher dose, the sponsor would want to use only the lower dose
- Mission
  - At interim analyses decide whether there is a substantial improvement with higher dose
  - Preserve inference for remaining does
- Issues
  - Effect of dose response on inference about remaining dose
  - Safety profile

217

Example 2: History

- Industry sponsored trials of new biologic agent in the treatment of sepsis
  - Investigation proceeded through Phase I, Phase II testing
  - Preliminary evidence of efficacy supported progressing to Phase III trial
  - In Phase III trial, some suggestion that treatment is most beneficial in patients with major organ failure

218

Example 2 : History

- Sponsor designing next Phase III (placebo controlled) trial
- Choices
  - Ignore the results from the subgroup analysis
  - Use prior entry criteria but terminate study early (and design new trial) if subgroups differ
  - Restrict patient entry to those with major organ failure

219

Example 2 : History

- FDA concerns
  - If next trial is conducted in different patient population, can this really be regarded as a confirmatory trial?

220

Example 2 : History

- Mission: Design an adaptive trial that will address potential differences within subgroups
  - Trial initially accrues patients per previous definitions
  - Predefined subgroups examined for consistency of effect at interim analyses
  - Subgroup may be dropped if no evidence of effect
    - (in agreement with previous trial)

221

Options

- "Adaptive Designs" allowing unlimited alterations
  - Fisher's self designing trial: Weighted Z statistics
  - Combining P values
  - Control of type I error will increase predictive value of positive going from phase II to phase III
- Prespecified group sequential rules
  - Allow determination of sampling distribution for a sufficient statistic
  - Addresses needs for estimation of effects
    - Though still problematic

222

Adaptive Testing of Subgroups

- Use up to three group sequential stopping rules
  - Test of treatment effect in combined group
    - efficacy, futility
  - Test for similar effect across subgroups
    - if not, stop subgroup B
  - If subgroup B previously stopped, test for efficacy, futility in subgroup A

223

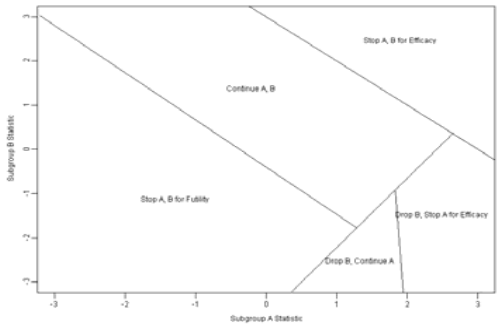
Adaptive Testing of Subgroups

- Alternative representation as partition of outcomes
- space for Z_{Aj} and Z_{Bj}
  - Each of the group sequential tests define lines demarcating regions for
    - stopping both subgroups simultaneously
    - stopping subgroup B for futility
    - stopping subgroup A for futility or efficacy
    - continuing to the next analysis

224

Adaptive Testing of Subgroups

- Alternative representation as partition of outcomes
- space for Z_{Aj} and Z_{Bj}



225

Adaptive Testing of Subgroups

- Issues to be considered
  - Hierarchy of tests
    - Stop both subgroups for efficacy
    - Stop subgroup B for lack of efficacy
      - if so, consider subgroup A efficacy, futility
    - Stop both subgroups for futility
    - Continue to next analysis

226

Adaptive Testing of Subgroups

- Issues to be considered (cont.)
  - Criteria for dropping subgroup B
    - Difference between subgroups (presence of interaction)
      - maintain combined group as long as possible
    - Limited effect in subgroup B
      - absolute criterion
    - Combination of the two

227

Adaptive Testing of Subgroups

- Issues to be considered (cont.)
  - Operating characteristics to control
    - Experimentwise Type I error
      - probability of declaring any effect when treatment ineffective in both subgroups
    - Probability of correctly identifying a beneficial effect in subgroup A
    - Probability of incorrectly identifying a beneficial effect in subgroup B
    - Probability of incorrectly dropping subgroup B

228

Adaptive Testing of Subgroups

- Issues to be considered (cont.)
  - Tradeoffs in presence of differential beneficial effects
    - Greater power if restricted to subgroup with most beneficial effect
      - (but may take longer to accrue)
    - Decreased indication if only approved in one of the subgroups
      - (but after earning money can restudy other subgroup at leisure)

229

Adaptive Testing of Subgroups

- Issues to be considered (cont.)
  - Sample sizes to accrue if subgroup B dropped
  - Parameterization of decision rules

230

Adaptive Testing of Subgroups

- Basic strategy explored here
  - Level α_1 group sequential test of combined subgroups
  - Level α_2 group sequential test to decide whether to drop subgroup B
    - based on differential effect, or
    - based on absolute effect in subgroup B
    - (α_2 only used as parameterization, though it should indicate tradeoff between combined and subgroup analysis)

231

Adaptive Testing of Subgroups

- Basic strategy explored here (cont.)
  - Level α_1 group sequential test of subgroup A if subgroup B dropped
  - α_2 chosen arbitrarily
  - α_1 chosen to achieve experimentwise level .025 test
  - If subgroup B dropped, subgroup A accrues to sample size planned for combined groups

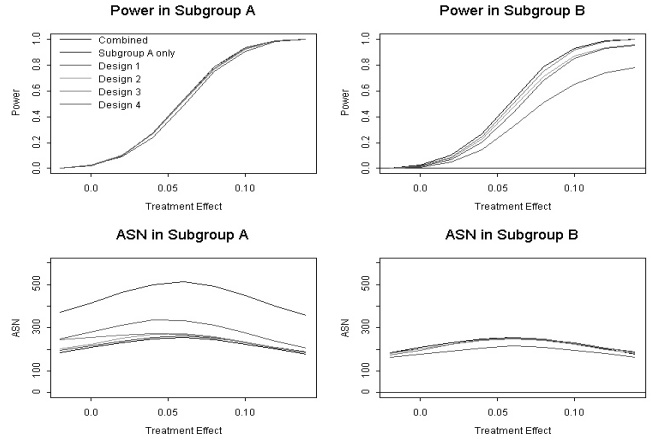
232

Adaptive Testing of Subgroups

- Comparisons
  - Nonadaptive trial with combined subgroups
  - Nonadaptive trial with subgroup A only
  - Four designs
    - Design 1 vs 2: early conservatism of subgroup
    - Design 1 vs 3: relative vs absolute criteria for dropping subgroup B
    - Design 1 vs 4: size of test used to drop subgroup B

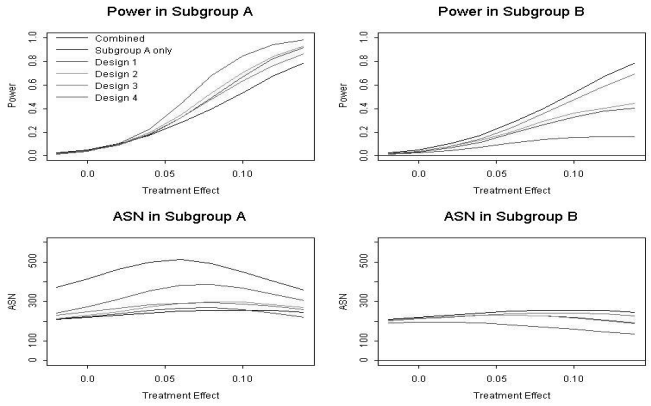
233

Equal Treatment Effects in Subgroups



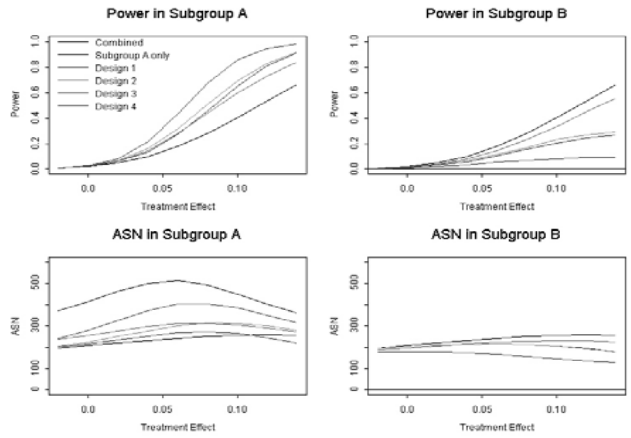
234

Small Treatment Effect in Subgroup B



235

No Treatment Effect in Subgroup B



236

Interpretation

- Role of adaptive trial as confirmatory
  - In absence of subgroup analyses, assumption is that same effect holds in both subgroups
  - Restriction to subgroup only decreases patients who might be exposed to ineffective treatment
  - If same ordering of subgroups observed in next trial, there is no reason not to regard results as confirmatory

237

Summary

- Post hoc identification of subgroups is always suspect, however
  - there is a range of operating characteristics that can be achieved by varying the decision rules even within the limited examples shown here
  - finding optimal rules should consider the costs and benefits of reduced indicaton versus reduced power
  - there is relatively little cost scientifically in considering the subgroups

238

Final Comments

239

Major Conclusions

- There is no substitute for planning a study in advance
  - At Phase II, adaptive designs may be useful to better control parameters leading to Phase III
  - At Phase III, there seems little to be gained from adaptive trials
    - We need to be able to do inference, and adaptive trials can lead to some very perplexing estimation methods
- "Opportunity is missed by most people because it is dressed in overalls and looks like work." -- Thomas Edison
- In clinical science, it is the steady, incremental steps that are likely to have the greatest impact.

240